1.1 Where CMOS is Going: Trendy Hype vs. Real Technology

Tze-Chiang (T.C.) Chen

IBM Fellow, Vice-President of Science and Technology, T.J. Watson Research Center, IBM Research Division, Yorktown Heights, NY

The development of silicon technology has been, and will continue to be, driven by system needs. Traditionally, these needs have been satisfied by the increase in transistor density and performance, as suggested by "'Moore's Law" [1] and guided by CMOS scaling theory [2]. As the silicon industry moves towards the 45nm node and beyond, two of the most important challenges facing Moore's Law and continued CMOS scaling are the growing standby power dissipation and the increasing variability in device characteristics. Actually, these effects are the embodiments of CMOS approaching atomistic and quantum-mechanical physics boundaries. However, the infusion of new materials and device structures will continue to extend CMOS device performance for a long time to come. Cooperative circuit/technology co-design, and architectures developed concurrently with these new device innovations will provide a comprehensive solution to the challenges of deep submicron CMOS.

I. INTRODUCTION

CMOS' approach towards atomistic and quantum-mechanical boundaries has motivated some pundits to profile devices based on nanotechnology, bio-electronics, or quantum computing as urgently-needed CMOS replacements which must be developed and made immediately available for general design use. In practice, these alternatives need many years for their support infrastructure to be developed; they are also not yet desperately needed. New materials coupled with innovations in circuit design and system architecture ensure at least another decade of continued CMOS development. Power dissipation and increasing variability emerge as first order concerns which must be addressed effectively in process development as well as in circuit design. How effectively these concerns are addressed depends on how well designers and process developers work together.

The history of CMOS development is marked by a series of challenges which have been overcome by ingenuity and hard work. In the first 35 years of process development, MOSFET scaling efforts were focused on extending performance, both by improving device speed as well as integrating more devices and functionality on chip. In the last 5 years, chip power and power density have been a major challenge. Contributors to the power problem include increased device density and device parametric variation, rising subthreshold leakage current and gate tunneling current and elevated device temperatures. Delay variation, induced by spatial and temporal process parameter tolerance, and voltage and temperature variation profoundly challenges timing precision. The inability to reduce gate insulator thickness prevents further channel length reductions, and leads to this crisis in the control of common device phenomena such as static leakage, short channel effect, and drain-voltage-induced barrier lowering. Inevitably, continued growth of active and static power arises from this inability to scale. Fortunately, help is on the way: the introduction of high-k gate dielectrics and metal gates in the near future will reduce gate insulator tunneling by orders of magnitude, and renew device length scaling. Feature size tolerance issues will become the ultimate limit to scaling. At roughly 2.5 years per CMOS generation introduction, these innovations will extend CMOS lifetime by approximately 10 more years, to approximately the 25nm node.

As CMOS approaches the 25nm node, stochastic threshold variation caused by dopant implant position in ultra-small inversion regions will give rise to more than 100mV of threshold variation. In addition, process proximity effects induced by layout, loading effects caused by device density, and gate line-edge roughness will bring additional contributions to variation. At the technology level, over the past 5 years, Optical Proximity Correction (OPC) and Reticle Enhancement Technologies (RET) have attempted to mitigate these effects, but in the future, closer cooperation between the process engineer, the circuit designer, and the EDA developer will be needed to integrate deeper into the technology what OPC and RET initiated. At the circuit level, emerging issues such as distortion in analog circuits (e.g. Phase-Locked Loops), and instability and parameter tracking in bi-stable circuits (e.g. latches and SRAM cells) require much more active intervention. At the architecture level, initiatives such as self-healing systems, self-biasing substrates, and simultaneous circuit/device diagnostics will extend the circuit's ability to survive and function given a wider range of sensitivities.

In short, cooperative circuit and technology co-design, and architectures developed concurrently with new device innovations will provide a comprehensive solution to the challenges of deep submicron CMOS. Developed independently, these solutions are disjoint, and improvements are only incremental. Developed together, these solutions assure that this last era in CMOS will be marked by even greater achievements in high performance computing solutions.

II. MITIGATING THE CMOS POWER CRISIS

Detailed power maps of fully operational microprocessors can be obtained with resolution far beyond that provided by on-chip thermal sensors. Specifically, SIMP (Spatially resolved Imaging of Microprocessor Power) images the temperature-dependent infra-red (IR) emission while the chip is cooled by a custom designed IR transparent heat sink [3]. Measurements of the thermal distribution of a fully operational IBM microprocessor during bootup of an operating system (2GHz, Vdd=1.4V), with average power 50W/cm^2,revealed local peaks, or "hot spots", far exceeding 100W/cm^2.

Dissipating such heat from high power chips is a challenge. Figure 1.1.1 shows cooling paths for both air- and water-cooled chips and typical thermal resistances. For air cooling, the thermal interface material (TIM) requirements are quite severe since TIMs must provide good thermal contact while accommodating the different thermal expansion coefficients of the package materials. The left bar in Fig. 1.1.1, representing a package with a good paste TIM, shows that the TIM thermal resistance limits the overall cooling capability of the package. With greatly improved TIMs, air cooling toward 100W/cm^2 under ideal conditions, as indicated by the middle bar, is possible. To handle even higher power densities and/or to further lower the junction temperature, liquid cooling may be employed. The leading technologies are microchannel cooling, jet impingement, and spray cooling. For example, cooling of 300 to 400W/cm^2 has been demonstrated with separable, single-phase Silicon microchannel coolers. These may be tested separately and attached to the chip with an advanced TIM which does not need to accommodate thermal expansion mismatch.

Reduced power dissipation density per device, with constant cooling capability and/or reduced voltage, can be traded for more devices on the chip. These additional devices can increase parallelism to recapture performance lost to lower frequency from diminished overdrive. A key to making the above scenario viable is for the designer to advance circuit topologies and processor architectures which reduce active power and logical vector activity.

With Dennard scaling, power density ideally remains constant. In practice this is no longer valid, as gate oxide thickness scaling has slowed. Figure 1.1.2 shows a cross section of a production

 1-4244-0079-1/06/$20.00

FET with oxide thickness of approximately 1nm. The high direct tunneling current through this oxide leads to increasing gate leakage contributions to standby current. Oxide thickness non-scaling further limits voltage scaling, and leads to worsened short channel effect. Figure 1.1.3 shows the effect of increasing gate current (at longer channel length) and increasing sub-threshold leakage at short channel length. Passive power is now comparable to active power [4]. Even with passive power still maintained as a fraction of total power, active power has increased due to the non-scaling of the power supply voltage. Total chip power for CMOS technology is now comparable to that seen in Bipolar technology of the early 1990s [5]. This power crisis is threatening the entire industry. To reduce passive power, the tunneling components must be reduced. Increasing active power further hinders any increase in clock frequency: to reduce active power it is necessary to reduce supply voltage V_{dd}.

II-A. Transistor Leakage Current Reduction

Three components comprise the majority of the contemporary MOSFET's static current: gate tunneling, band to band tunneling current, and sub-threshold leakage, all of which have grown by greater than the technology scale factor. Promising high-k gate dielectric materials provide the capacitive coupling of a thin dielectric with the tunneling characteristics of a thick dielectric. The gate tunneling current will be significantly reduced by incorporating metal gate and high-k materials into the MOSFET. Subthreshold currents will be suppressed by controlling the mean and standard deviation of MOSFET threshold voltage distribution. Managing threshold variability and introducing back-gated MOSFETs are two approaches which will mitigate these currents.

II-A-1. High-K/Metal Gate – Mitigating Oxide Tunneling Current

Advanced gate dielectric films with dielectric constants higher than that of SiON (k>5.5) have been actively explored in the semiconductor industry and in academia for the past 15 years. The requirements for advanced dielectrics to replace traditional SiON include i) thermal stability, ii) high carrier mobility, iii) high permittivity, iv) good tunneling reduction, v) high reliability and vi) compatibility with metal gate electrodes and the underlying silicon substrate. The challenge is achieving these properties while maintaining compatibility with standard CMOS fabrication processes. The industry has settled on Hf-based gate dielectrics which include pure HfO_2 with dielectric constant of 20 and compounds such as HfSiAlO, HfSiOx, and HfSiON with dielectric constants of 12 to 15. These materials provide high mobility and have sufficient thermal stability and significant tunneling reduction compared to SiON.

Figure 1.1.4 shows the reduction in the inversion thickness (Tinv) and the silicon dioxide equivalent thickness (ToxGL) from the 130nm to 45nm nodes. The Tinv thickness is scaled as the Lpoly decreases in each successive node. Inversion thickness scaling is achieved by either reducing the physical thickness or increasing the dielectric constant. As the gate dielectrics are thinned, the corresponding gate tunneling increases exponentially. As the dielectric film has scaled from Tinv of 23 Angstroms to 17 Angstroms the gate leakage has increased from approximately 1A/cm^2 to several hundred A/cm^2. In the latest technology node, deviation from continued inversion thickness scaling to avoid this tunneling is noteworthy; unanswered, it will result in loss of gate channel control. The lack of scaling in the 65nm node is driven primarily by excessive gate leakage and reliability. Figure 1.1.4 shows future Tinv and gate leakage reduction exhibited by high-k and metal gates. The introduction of these materials allows our industry to continue electrical thickness scaling.

The cross-sectional TEM image in Fig. 1.1.2 shows a dramatic increase in physical thickness with high-k materials that can provide several orders of leakage reduction, with the same or even improved inversion thickness [6]. The emergence of high-k materials will coincide with the simultaneous introduction of band-edge metal gate electrodes. Polysilicon gate electrodes are unsuitable in high-k devices due to polysilicon depletion effects which increase the inversion thickness, and band-alignment issues which lead to excessively high threshold voltages [7]. A significant constraint is the need for separate band-edge pFET and nFET metals gates for improved short channel and drain-induced barrier lowering control. This requires a steep materials and process learning curve, since metals-on-Hf-based oxide work-functions differ significantly from those reported on SiO_2 [8]. Recent advances establishing compatible metal gates and high-k materials have made their introduction into the standard CMOS process highly likely within the next three years.

As shown in the high-k curve of Fig. 1.1.3, high-k dielectrics return CMOS static power to that of classic scaling by mitigating the tunneling current, and by improving channel control. The lower capacitance of shorter channels enabled by better short channel control (resulting from lower effective oxide thickness or "EOT") and lower V_{dd} reduce active power as well.

II-A-2. Threshold Voltage Control

High performance microprocessor chips have over 100 million MOS transistors. Due to manufacturing process control limitations, transistor characteristics can vary within a chip and from chip to chip. In particular, threshold voltage variation arising from gate length variation significantly impacts chip power and performance. Minimizing threshold voltage variations will deliver a superior power/performance tradeoff. Figure 1.1.5 shows simulation results for a 90nm CMOS technology. Circuit performance is estimated by 20 stages of FO4 inverter delay time. Energy per computation is calculated by dividing total power by clock frequency for an FO4 inverter with nFET and pFET width of 0.7μm and 1.4μm respectively. Three cases of threshold voltage tolerance (defined as the difference between nominal and minimum threshold voltages) are examined: 0V (ideal), 50mV (typical) and 100mV (worse case). For the ideal case, supply voltage (V_{dd}) and threshold voltage (V_T) are optimized for minimum power at each performance level using the methodology outlined in [9]. For practical cases, the nominal V_T is raised over optimized V_T to accommodate the V_T tolerance. The result shows a steep penalty for poor V_T tolerances. A 100mV V_T tolerance translates to a 100% increase in energy per computation at the same performance level of 2GHz, or a 25% decrease in performance at the same energy of 3fJ per clock cycle. Threshold voltage tolerance due to random dopant distribution and gate length variation will be discussed in section III. Undoped channel design, either ground plane devices or ultra-thin body devices on SOI (UTSOI), can help to minimize dopant fluctuation-related V_T variation. Back bias techniques, either well bias in a bulk device, or back gate bias for thin body SOI devices, can be used to adjust V_T locally to compensate for across-chip and chip-to-chip V_T variations, and will be discussed in section IV.

II-B Supply Voltage Reduction

The second means of addressing power consumption is to reduce operating voltage [9]. Supply voltage reduction quadratically mitigates active as well as static power. Further, reducing V_{dd} significantly reduces gate insulator tunneling for a fixed dielectric thickness. But because overdrive has become precious in deep submicron CMOS, operating at reduced voltages requires adding complexity in supply distribution and modulation. Performance in the new CMOS, however, must be measured in total system

throughput rather than in GHz. Figure 1.1.6 demonstrates the dramatic efficiency improvement that can be realized by operating at reduced V_{dd}. It is ironic to note that, just as with CMOS' replacement of HBTs, a lower performance / lower power technology ultimately will deliver superior system throughput because of the higher integration it enables. Blue Gene-L, the world's fastest supercomputer, uses processors running at 700MHz - substantially slower and cooler than high performance 2 to 3GHz workstation uni-processors.

II-C Architecture and Design Optimization

Device design, *supply voltage reduction*, and *tolerance control* discussed above have limited ability to resolve the power crisis. The rest must be accomplished by circuit designers and microprocessor architects. Not only must circuits and architectures assume more power and variability management responsibility; they must improve throughput without introducing more complexity. For this reason, static combinatorial CMOS remains the topology of choice in high performance complex state machines. As pressure for performance increases, more efficient structures such as pass-gate-based logic become more interesting [10].

Transaction rates have been improved by increasing the processor pipeline depth. Deeper "pipes" more fully utilize compute resources, but reduce the amount of work done per cycle; with fewer stages per cycle, deeper-pipelined machines run at higher frequency. Figure 1.1.7 shows an IBM analysis looking for the optimum number of INVFO4 equivalent stages per cycle [11]. As the number of INVFO4 delays per cycle drops, the frequency increases and the number of instructions per cycle (IPC) decreases. Since fewer transactions are completed per cycle, more registers need to be added and they are clocked at higher frequency, causing the power to go inversely with FO4 depth. The relative performance shows a broad optimum.

Given recent power restrictions, the architect is motivated to "go long" and park the design to the right of the arrow in Fig. 1.1.7. For the same amount of function, the "longer-stages/cycle" machine will use less power and have better timing precision. Note that register count (and power) increases monotonically with the pipeline depth. It follows that the global chip designer decides area and power as well as performance when selecting the architecture of a new machine. In a power constrained environment, those choices make big differences and vary from preferences of even only one generation ago.

II-D 3D Silicon

If power dissipation has been managed to a sufficiently low level via tolerance control, static leakage current mitigation, and effective low power circuit topologies, then implicitly heat dissipation can be accommodated with package technology. The reward of these efforts is then that higher levels of performance per unit volume can be realized through the use of three-dimensional integrated circuits, which stack active device layers in the vertical direction. Such three-dimensional integrated circuits (3D-ICs) are expected to exhibit significant performance and power advantages over their 2D counterparts [12]. Many of these advantages are derived from the reduced wire-length distributions for circuits designed in stacked layers, which in turn lead to lower overall latencies and power dissipation for such designs. Significant benefits can also come from architectural advantages of ultra-high bandwidth, enabled between the layers. The ability to integrate disparate process technologies, optimized for individual device layers, is an additional benefit. IBM is exploring 3D-integration technologies for both bulk and SOI substrates. Our SOI 3D-IC fabrication-process flow [13] is depicted in Fig. 1.1.8. In this technology, one circuit layer is transferred from its original substrate onto the top of another circuit.

III. VARIABILITY

Chip variability is of great importance, influencing yield and overall chip-level performance. Variability manifests itself in a multitude of processes and can be global (chip to chip, wafer to wafer, etc.) or local (inherent to the circuit instance). Advanced process controls (APC) (i.e. run-by-run control, fault detection and classification) can effectively address chip-to-chip, wafer-to-wafer and lot-to-lot mean variations by minimizing global variation. *Regional systematic variations* across the wafer and across the chip are typically corrected by process modifications. ACLV, for example, can be partially corrected by judiciously adjusting the exposure dose across the wafer during the step-and-scan wafer exposure. *Regional yet nonsystematic in-chip variation* cannot be addressed by APC, process design or process modification; it requires a far more advanced solution. In the future, self-correcting, autonomic circuits will be required to address this type of variability. These circuits rely upon on-chip monitors to continuously test function and performance, self-correcting by turning on and off peripheral circuits to adapt. *Local in-chip variations* resulting from processing pattern density effects require process design modifications. Examples include optical proximity correction (OPC) in advanced lithography to minimize across-chip linewidth variation (ACLV). Other examples include the addition of pattern fill to improve copper wire planarization and shallow-trench-isolation divot reduction. Thermally-induced variation during rapid-annealing is the more recent of these effects and can be mitigated through the thin film design. *Highly localized in-chip variations* due to the fundamental, underlying physics or chemistry of some processes cannot be resolved by the aforementioned techniques.

Stochastic ion-implanted dopant variation has no known remedy; it is simply inherent to the process and cannot be mitigated. Figure 1.1.9 illustrates the nature of random dopant placement and its effect. As transistors are scaled, fewer dopant atoms per device results in larger variation in threshold voltage. The larger threshold voltage mismatch within SRAM cells as the cell size decreases will seriously impact circuit yield at the 45nm node. Another example of highly localized in-chip variation shown in Fig. 1.1.9 is the line edge roughness (LER) effect [14]. As device critical dimensions (CDs) continue scaling into sub-50 nm regime, a growing contribution to performance variation comes from the deviation from nominal feature size, referred to as line width roughness (LWR). LWR is defined as the 3σ line width deviation measured at ″10nm intervals, over a line length of ≥2μm. A fundamental understanding of the sources and impact of LER and LWR, as well as it's characterization, correlation to performance, and it's reduction have become semiconductor manufacturer priorities. Many researchers have explored the sources of this variation; it is often largely attributed to patterning steps, lithography and reactive ion etch (RIE) [15].

III-A. Relationship between Variability and Design

The three major challenges in dealing with random variability are characterization, reduction, and accommodation. Design can play a vital role in addressing all three.

Characterization of variability is frequently thought of primarily as a DC parametric exercise, directly measuring distributions of quantities such as V_T, Id, and Lp. At the product level, the primary source of information on variability comes from diagnosis of the product itself or from high speed test structures embedded within the product chip. The design of appropriate on-board high speed test structures on dedicated test chips, on the product kerf, and within the product itself is an important aspect of variability characterization. Carefully crafted ring oscillators and DC testable, self-timed, pulse-based test structures are examples of structures that are proving to be valuable sources of information

on variability [16]. In addition appropriate test structures embedded in the chip design itself can be used to directly evaluate variability within the product and even to provide inputs for decisions on dynamic change in the operation of the chip [17].

Design can play an important role in variability reduction. Local variations in layout often lead to significant variations in device parameters through their impact on lithography, etch, CMP, and local stress. Future generations of technology will require an increased focus of design for manufacturing (DFM), particularly as variability in device performance becomes an even more serious problem. DFM can optimize layout robustness against both random and systematic process yield detractors [18]. DFM from the lithography perspective is the means by which design layouts respect specific restrictions so that they robustly print. Examples of such restricted design rules (RDRs) include (a) the avoidance of lithographically forbidden pitches; (b) single orientation of narrow gates; (c) restrictions on the number of linewidths that are allowed, and (d) placement of narrow features on a quantized grid so that a uniform proximity environment is attained. RDRs can also afford reduced CD variation such as across chip or field linewidth variation (ACLV, AFLV) that can manifest in device performance variation.

Enormous opportunities exist to accommodate variability through design practices. In the absence of variability, the performance of a design could be predicted to the limit of model's accuracy. In the presence of variability, the situation rapidly becomes more complex. Statistical techniques are now often used to establish a likely distribution for the projected power-performance [19]. The accuracy of this approach depends directly on knowledge of the nature of the variability. Correspondingly, the sensitivity of a design to variability can be reduced by leveraging that same knowledge. Logic depth of paths within a processor forms a good example. Recent microprocessor architectures extend the trend to reduce the number of logical stages in a path. If there is a strong random component in the delay of logic gates, then the variability in the path delay worsens as the path is shortened. Increased variability eventually becomes great enough to compromise gains arising from shortening paths. Another design option for dealing with variability is a dynamic approach in which appropriate test structures embedded in the chip design are continuously monitored during the operation of the chip [17]. Feedback is provided to the power supply voltage or back-bias to accommodate initial variations across the chip and to compensate for workload variations, while remaining within allowed power/performance limits [17].

Analog circuits are particularly affected by variability, impacting linearity, VCO tuning range, signal-to-noise margin, and the accuracy of circuit models. As the supply voltage decreases, the reduced headroom further magnifies existing variation in threshold voltage. To counter this problem, the ITRS roadmap calls for analog circuits to continue to require higher V_{dd} compared to digital circuits. This implies the use of separate technology elements such as multiple gate oxide thicknesses and additional circuitry such as level shifters to interface circuits with different supply voltage. The increasing difficulty of analog circuit design will encourage replacement of analog functionality by digital signal processing built in static CMOS digital circuitry [20].

IV. ENHANCING DEVICE PERFORMANCE

The International Roadmap for Semiconductor (ITRS) projects that MOSFETs with equivalent oxide thickness of 5A and junction depths less than 10nm will be in production in the next decade [21]. While 6nm gate lengths MOSFETs have been demonstrated [22], performance and manufacturability problems remain. High-k/metal gates to mitigating gate leakage, alternative doping techniques for shallower more abrupt junctions, and alternative device structures will likely be needed for sub-30nm devices.

Although ever more challenging to increase CMOS device performance, innovative CMOS device designers continue to be successful. A number of techniques have enhanced CMOS device performance appreciably. These include strain-induced mobility enhancement, using silicon in unusual crystal orientations, and creating novel device structures, to improve the electrostatics of the devices.

IV-A. Mobility Enhancement

Mobility enhancement techniques are attractive for performance enhancements beyond those derived from device scaling alone. Straining the silicon and building n-type and p-type MOSFETs on different crystal orientations are promising methods for mobility enhancement. Biaxial tensile strained wafers [23] can introduce strain on the initial substrate. Strained-silicon directly on insulator (SSDOI) structures [24] eliminates the SiGe layer before transistor fabrication, thereby providing higher mobility while mitigating the SiGe-induced material and process integration problems [24, 25].

Process techniques can also introduce strain. The contact etch stop nitride layer can act as the stress film [26]. Dual stress liner (DSL) [27], with compressive stress for pFET and tensile stress for nFET, has been implemented in IBM's 90nm technology node. Higher stress can be generated with thicker films. Local strain can also be applied to the channel through the Stress Memorization Technique (SMT) [28]. Epitaxially-grown strained SiGe (e-SiGe) can be embedded in the S/D regions [29] and extension location [30]. When the lattice spacing of this SiGe material is larger than Si, uni-axial compressive strain is induced in the channel. This structure can produce significant pFET hole mobility improvement. The compressive stress is mainly dependent on the e-SiGe thickness and percent Ge content.

Inversion layer mobility depends on surface orientation and channel direction. The hole mobility is more than two times higher on (110) surface while electron mobility is the highest on the standard (100) surface. Hybrid orientation technology (HOT) [31] using pFETs on (110) surface orientation and nFETs on (100) orientation is promising for future CMOS applications because the fabrication processes are fully compatible with current VLSI technology; no new materials are introduced for the pFET performance improvement. Device and circuit performance enhancement using the HOT technology have been reported [31]. The most popular technologies for mobility enhancement are summarized in Fig 1.1.11.

IV-B. Novel Devices

Threshold voltage variations can significantly degrade system performance. An additional terminal can control the device threshold voltage. Subthreshold leakage control by body biasing [32, 33] in bulk technology has been proposed to manage leakage power. Back-gate MOSFETs with thin buried oxide (BOX) is a superior solution over bulk body biasing [34]. Ultrathin SOI MOSFETs effectively reduce short-channel-effects and eliminate leakage paths [22, 35]. The silicon thickness for short-channel effect control can be relaxed by using a more complex "double-gate" structure offering improved electrostatic control. Scalability for double-gate FET improves by a factor of 2.5 to 3 [35]. Because the double-gate device operates at much lower vertical electric fields, mobilities are higher [34]. The FinFET [36, 37] is the simplest double gate structure [38] to implement.

IV-C. Parasitic Resistance

Parasitic resistance is becoming increasingly important in the nanometer regime. Parasitic resistance arises from S/D extension

resistance, contact resistance at the silicide/silicon interface, and contact via resistance. Fast anneal with activation beyond solid-solubility can mitigate S/D extension sheet resistance. Nickel-silicide supports line width scaling and improves contact resistance. Lower barrier height Silicides [39] may be needed beyond 45nm node, however. The contact via resistance is perhaps the most problematic. Figure 1.1.10 shows the contact via dimensions and resistance trend as function of technology node.

V. CONCLUSION

The future for our industry contains a *challenge*, an *agreement*, and an *opportunity*. The *challenge* to the circuit design community is to develop circuits and architectures which deliver high throughput computing at low power. We know it is possible; the biological wet processor sets the precedent for high throughput, high parallelism, at very low power. If the circuit design community can deliver on this challenge, the process development community *agrees* to provide materials and solutions such as high-k dielectrics, three-dimensional chips, and stable low voltage MOSFETs to enable new design points. Equipped with these *opportunities*, future product design points will continue to move the state of the art forward into yet more astonishing capabilities.

Acknowledgements:
The author acknowledges contributions from members of the Science and Technology Department at IBM Research, in particular Tak Ning, Kerry Bernstein, James Stathis, Ghavam Shahidi, Dave Medeiros, Mark Ketchen, Mekei Ieong, and James Warnock for their valuable discussions and suggestions.

References:

[1] G. E. Moore, "Cramming more Components on to Integrated Circuits," *Electronics*, vol. 38, no. 8, April 19, 1965; "Progress in Digital Integrated Electronics," *IEDM Tech Dig*., pp. 11-13, 1975.
[2] R. Dennard et al., "Design of Ion-Implanted MOSFET's with Very Small Physical Dimensions," *IEEE J. Solid-State Circuits*, vol. SC-8, pp. 256-268, 1974.
[3] H. F. Hamann et al., "Spatially Resolved Imaging of Microprocessor Power," *ISSCC Dig. Tech. Papers*, p.16, Feb., 2005.
[4] E. J. Nowak, "Maintaining the Benefits of CMOS Scaling when Scaling Bogs Down," IBM J. Res. Dev. vol. 46, no. 2/3, p. 169, 2002.
[5] R. R. Schmidt and B. D. Notohardjono, "High-End Server Low-Temperature Cooling," *IBM J. Res. Dev.,* vol. 46, no. 6, p. 739 2002.
[6] V. Narayanan et al., Materials Research Society Spring Meeting, April, 2003.
[7] C. Hobbs et al., *IEEE Trans. Elec. Dev.*, vol. 51, pp. 971-977, 2004.
[8] H. Takeuchi et al. "Impact of Oxygen Vacancies on High-k Gate Stack Engineering," *IEDM Tech. Dig.,* pp. 829 - 832, 2004.
[9] J. Cai et al., "Supply Voltage Strategies for Minimizing the Power of CMOS Processors," *Symp. VLSI Tech*., p. 102, 2002.
[10] K. Yano et al., "Top-Down Pass-Transistor Logic Design," *IEEE J. Solid-State Circuits*, vol. 31, pp. 792-803, 1996.
[11] K. Bernstein, "Caution Flag Out: Microarchitecture's Race for Power Performance," *Proc. 36th Annual International Symposium on Microarchitecture*, 2003.
[12] K. W. Guarini, et al., "The Impact of Wafer-Level Layer Transfer on High Performance Devices and Circuits for 3D IC Fabrication," *in Proceedings of the 203rd Meeting of the Electrochemical Society, vol. PV* 2003-13, Paris, France, 2003.
[13] A. W. Topol, et al., "Enabling SOI-Based Assembly Technology for Three-Dimensional (3D) Integrated Circuits (ICs)," *IEDM Tech Dig.*, 2005.
[14] A. P. Mahorowala et al., "Impact of Thin Resist Processes on Post-Etch LER," *Proc. SPIE 5039*, 213, 2003.
[15] P. K. Montgomery et al., "Reduction of Line Edge Roughness and Post Resist Trim Pattern Collapse for Sub 60nm Gate Patterns Using Gas-Phase Resists Fluorination," *Proc. SPIE 1024*, 5753, 2005.
[16] M. Ketchen, M. Bhushan, and D. Pearson, "High Speed Test Structures for In-line Process Monitoring and Model Calibration," *Proc. 2005 IEEE Conf. Microelectron. Test Structures*, pp. 33-38, April 2005.
[17] C. Poirier et al., "Power and Temperature Control on a 90nm Itanium-Family Processor," *ISSCC Dig. Tech. Papers*, pp. 304-305. 2005.
[18] L. Liebmann et al., "Integrating DfM Components into a Cohesive Design-to-Silicon Solution," in Design and Process Integration for Microelectronic Manufacturing III, ed. L.W. Liebmann, *Proc. SPIE 5756*, pp. 1-12, 2005.
[19] C. Visweswariah et al., "First-Order Incremental Block-Based Statistical Timing Analysis," *DAC*, pp. 331-336, 2004.
[20] M. Oprysko, panel session, "Will Continued Process-Node Shrinks Kill High-Performance Analog Design?" *IEEE CICC*, Sept. 2005.
[21]http://www.itrs.net/Common/2004Update/2004Update.htm
[22] B. Doris et al., "Extreme Scaling with Ultra-Thin Si Channel MOSFETs," *IEDM Tech Dig*., pp. 267-270, 2002.
[23] K. Rim et al., "Mobility Enhancement in Strained Si NMOSFETs with HfO_2 Gate Dielectrics," *Symp. VLSI Tech.*, p. 12, 2002.
[24] K. Rim et al., "Fabrication and Mobility Characteristics of Ultra-Thin Strained Si Directly on Insulator (SSDOI) MOSFETs," *IEDM Tech Dig*., pp. 49-52, 2003.
[25] A. Thean et al., "Performance of Suptercritical Strained Silicon Directly on Insulator," *Symp. VLSI Tech*., p. 134, 2005.
[26] A. Shimizu et al., "Local Mechanical-Stress Control (LMC): A New Technique for CMOS-Performance Enhancement," *IEDM Tech. Dig.*, pp. 433-436, 2001.
[27] H. S. Yang et al., "Dual Stress Liner for High Performance sub-45nm Gate Length SOI CMOS Manufacturing," *IEDM Tech. Dig*., pp. 1075-78, 2004.
[28] D. V. Singh et al, "Stress Memorization in High-Performance FDSOI Devices with Ultra-Thin Silicon Channels and 12nm Gate Lengths," *IEDM,* 2005.
[29] T. Ghani et al., "A 90nm High Volume Manufacturing Logic Technology Featuring Novel 45nm Gate Length Strained Silicon CMOS Transistors," *IEDM Tech. Dig*., pp. 978-980, 2003.
[30] P. R. Chidambaram et al., "35% Drive Current Improvement from Recessed-SiGe Drain Extensions on 37nm Gate Length PMOS," *Symp. VLSI Tech*., pp. 48-49, 2004.
[31] M. Yang et al., "High Performance CMOS Fabricated on Hybrid Substrate with Different Crystal Orientations," *IEDM Tech Dig.*, pp. 453-456, 2003.
[32] S.-F. Huang et al., "Scalability and Biasing Strategy for CMOS with Active Well Bias," *Symp. VLSI Tech.*, pp. 107-108, 2001.
[33] S. Borkar, "Circuit Techniques for Subthreshold Leakage Avoidance, Control, and Tolerance," *IEDM Tech. Dig*., pp. 421-424, 2004.
[34] M. Ieong et al., "Experimental Evaluation of Carrier Transport and Device Design for Planar Symmetric / Asymmetric Double-Gate / Ground-plane CMOSFETs," *IEDM Tech. Dig*., pp. 441-444, 2001.
[35] M. Ieong et al., "Ultra-Thin Silicon Channel Single- and Double-Gate," *Ext. Abst. SSDM*, pp.136-137, 2002.
[36] J. Kedzierski et al., "High-Performance Symmetric-Gate and CMOS-Compatible Vt Asymmetric-Gate FinFET Devices," *IEDM Tech. Dig.*, pp. 437-440, 2001.
[37] D. Hisamoto et al., "A Fully Depleted Lean-channel Transistor (DELTA) — A novel vertical ultra thin SO1 MOSFET," *IEDM Tech. Dig.*, pp. 833-836, 1989.
[38] H.-S. P. Wong, "Beyond the Conventional Transistor," *IBM J. Res. Dev*., pp. 133-167, 2002.
[39] M. Ieong et al., "Comparison of Raised and Schottky Source/Drain MOSFETs Using a Novel Tunneling Contact Model," *IEDM Tech. Dig.*, pp. 733-736, 1998.

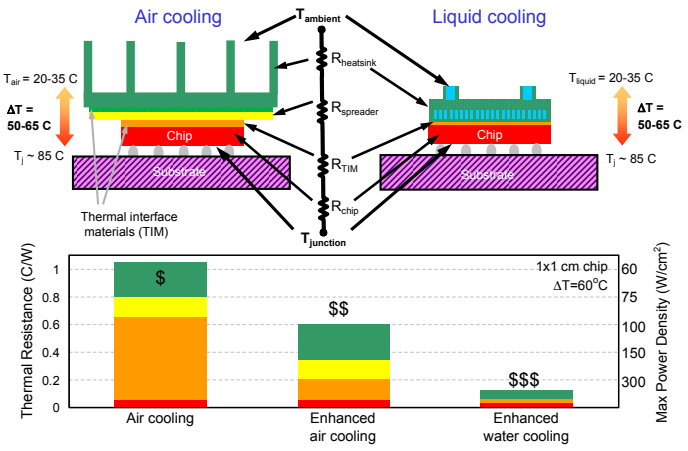

Figure 1.1.1: The bar segments (bottom to top) correspond to the chip, TIM layer, heat spreader, and heat sink or microchannel cooler.

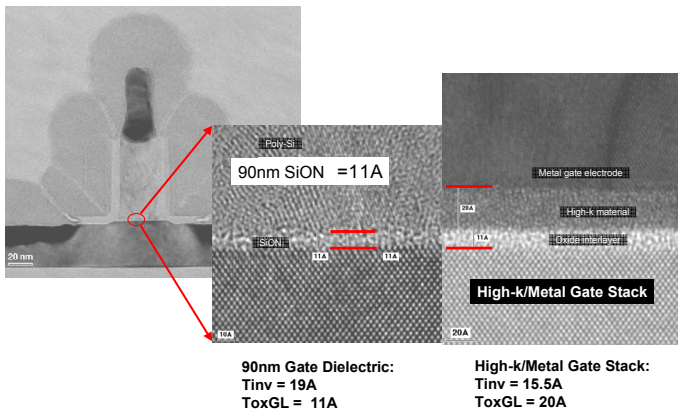

Figure 1.1.2: Gate stack cross-sections.

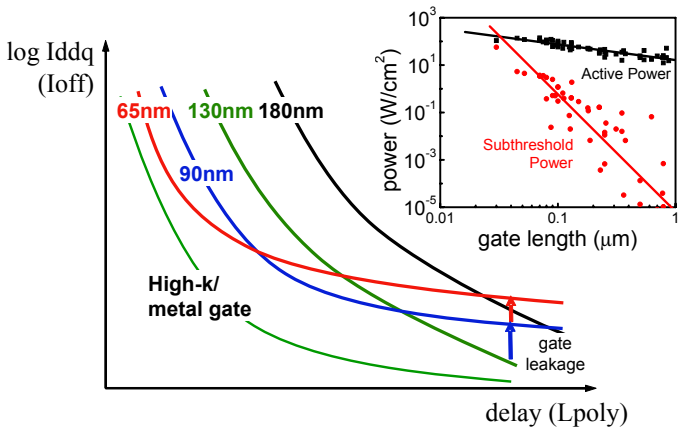

Figure 1.1.3: Technology comparisons. Inset: Comparison of passive and active power.

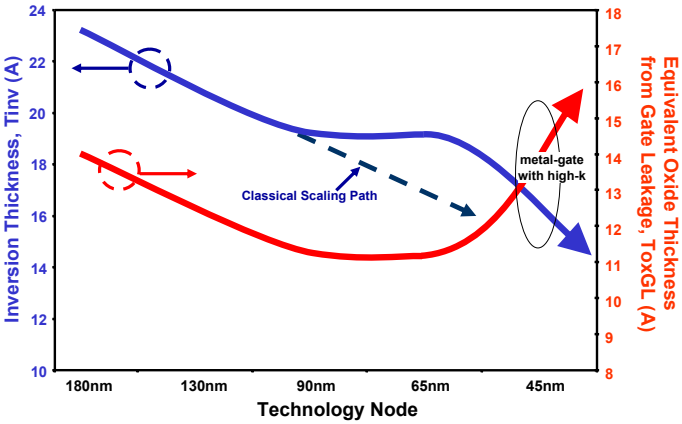

Figure 1.1.4: Oxide thickness scaling.

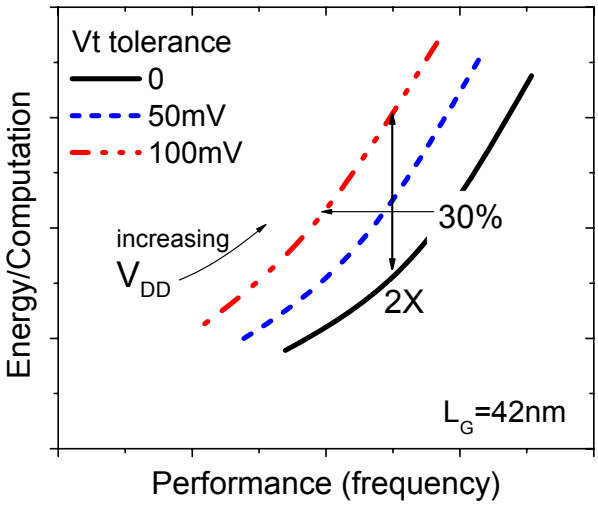

Figure 1.1.5: Effect of variability on performance.

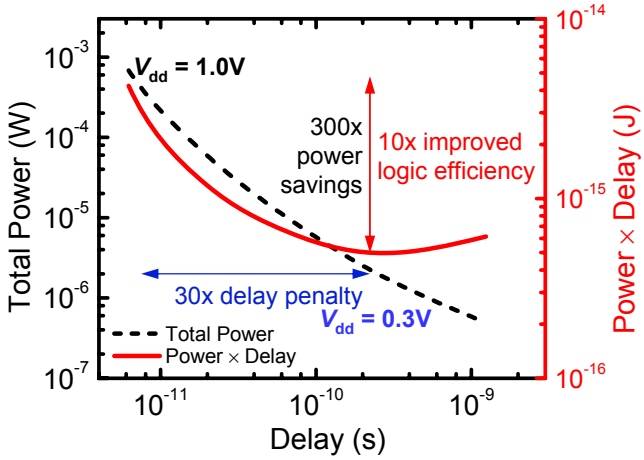

Figure 1.1.6: Experimental results for 35 nm ring-oscillator. Reduced V_{dd} can result in 10x improved logic efficiency.

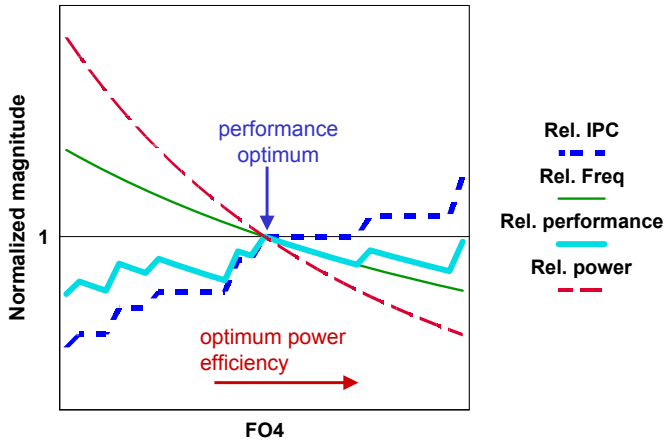

Figure 1.1.7: Optimum pipeline depth for a given microprocessor workload.

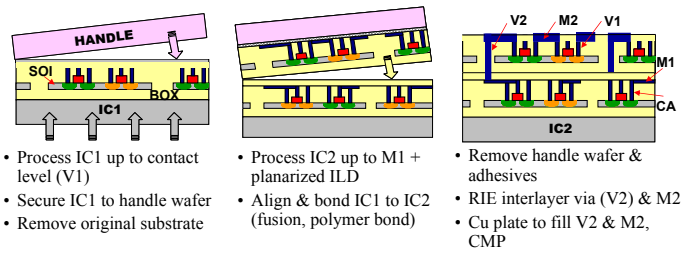

Figure 1.1.8: 3D-IC fabrication-process flow.

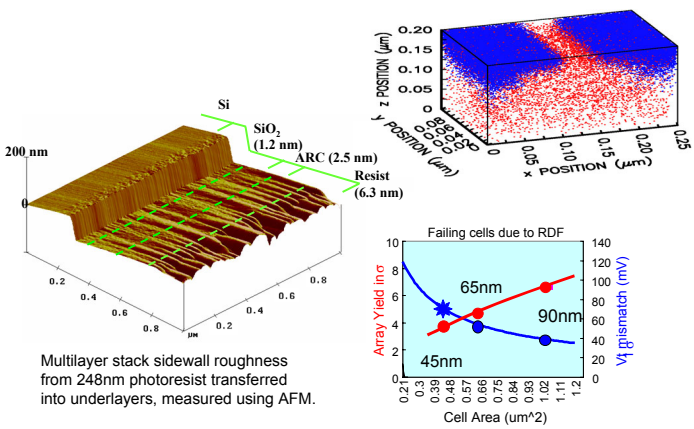

Figure 1.1.9: Sources and impact of variability.

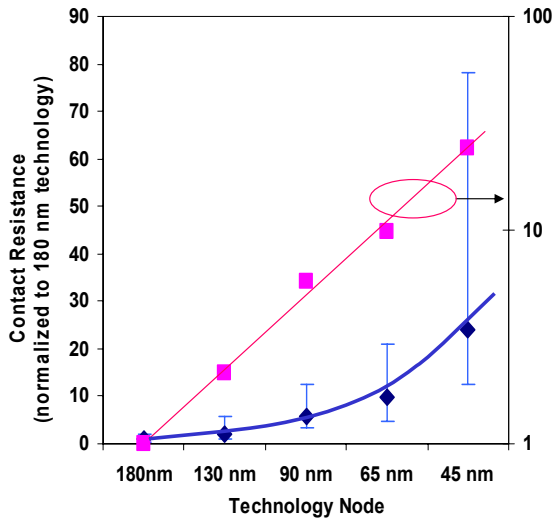

Figure 1.1.10: Historical trend in contact via resistance on linear and log scale.

	NMOS	PMOS
Biaxial Tensile Strain	↑	↑
Contact etch-stop liner (DSL)	↑	↑
SMT	↑	o
e-SiGe	o	↑
Substrate Orientation (HOT)	o	↑
Channel Orientation (<100>)	o	↑

Figure 1.1.11: Summary of feasible mobility enhancement techniques and their associated effect on NMOS and PMOS.